利用结构化 SVM 结合 CNN 的层次化目标检测与人体姿态估计方法 *

孙新领 [1]，张 皓 [1]，赵 丽 [2]
(1. 河南工学院 计算机科学与技术系，河南 新乡 453003; 2. 山西大学 软件学院，太原 030013)

摘 要：针对现有姿态估计方法不能准确提取特征参数的问题，提出了一种基于结构化支持向量机（SSVM）与卷积神经网络（CNN）的层次化模型。首先，展示了一个基于 PS 部件模型的 SSVM 如何实现为一个两层的神经网络，其中第一层是卷积层，另一层是损失增强推理层；通过将模型的结构化形式转换为模型中的一个神经网络，提出的方法可以同时学习结构模型和外观模型，然后反向传播误差以学习底层的可学习参数，这些参数可从外观模型特征中提取出来；最后，将 SSVM 模型转换为神经网络模型，将误差反向传播到较低层，并计算确切的 SSVM 损失，同时通过基于次梯度的方法来学习原始 SSVM。将该模型与当前较为先进的识别模型进行了对比，结果证明提出的层次化模型的识别成功率比对比方法平均高 6%，具有更强的识别性能。
关键词：人体姿态估计；外观模型；深度神经网络；卷积层；结构化 SVM
中图分类号：TP391.41 doi: 10.19734/j.issn.1001-3695.2018.11.0855

Hierarchical target detection and human body attitude estimation based on structured SVM and CNN

Sun Xinling[1], Zhang Hao[1], Zhao Li[2]
(1. *Dept. of Computer Science*, *Henan Institute of Technology*, *Henan Xinxiang* 453003, *China*; 2. *School of Software*, *Shanxi University*, *Taiyuan* 030013, *China*)

Abstract: Aiming at the problem that the existing attitude estimation method can not accurately extract the feature parameters, this paper proposed a hierarchical model based on structured support vector machine (SSVM) and convolutional neural network (CNN) . First, it showed how a SSVM based on the PS component model could be implemented as a two-layer neural network, where the first layer was the convolutional layer and the other layer was the loss-enhanced inference layer. Then, by transforming the structured form of the model into a neural network in the model, the proposed method could simultaneously learn the structural model and the appearance model, and then backpropagated the error to learn the underlying learnable parameters. These parameters could be derived from the appearance model features. Extracted out. Finally, the SSVM model was transformed into a neural network model, the error was propagated back to the lower layer, and the exact SSVM loss was calculated, while the original SSVM was learned by the sub-gradient-based method. Comparing the model with the current advanced recognition model, the results show that the proposed success rate of the hierarchical model is 6% higher than the comparison method and has stronger recognition performance.
Key words: human pose estimation; appearance model; deep neural network; convolutional layer; structured SVM

0 引言

目前，深度学习和特征学习是解决分类、检测等问题的流行方法。包括使用卷积神经网络（convolutional neural network，CNN）进行人脸检测[1,2]，使用深度神经网络(deep neural networks, DNN)进行行人检测[3]，使用 DNN 进行人体姿态估计（human pose estimate，HPE）[4]，使用受限玻尔兹曼机(restricted Boltzmann machine, RBM)进行人脸特征跟踪[5]以及使用深度学习对物体分割进行形状先验检测以及使用深度网络进行物体检测等。为了联合使用 latent SVM 和深度学习，通常是使用 DNN 提取特征，然后用于 latent SVM 的学习，构建分类器。

一些学者提出了独特的解决方法，例如，文献[6]提出了一种带有附加潜变量的图形结构树模型，精心设计了叶节点变体和潜在节点，它们控制叶节点的变化，而增加了一个用于推理的循环模型。文献[7]关注的是将部位聚类为多模态可分解模型。试图通过参数化几何变量来获得更好的先验模型，从而改进图像结构。但是如果模型得到改进，所有这些方法都必须学习结构模型参数。Latent SVM 是学习这些模型参数的标准方法。文献[8]在特征提取阶段从 CNN 中提取了一个金字塔特征，然后缓存提取的特征，再在第二阶段使用 latent SVM 进行学习；在第二阶段中，latent SVM 通过在 SVM 优化和推理组合优化之间切换来学习所有模型参数。然而，这种方法存在固有的问题，因为这种方法分为两个不同的阶段进行，它不能学习由深度学习特征而提取的参数。而且由可学习的特征所提取的参数不能基于 Latent SVM 的误差来更新。

基于上述分析，为了解决现有姿态估计方法不能准确提取特征参数的问题，提出基于结构化 SVM 卷积神经网络的层次化模型。部件模型是视觉识别中一种重要的结构化建模方法，特别是 DMP（deformable part model）和 PS（pictorial structure）模型[9,10]。DPM 一元过滤器方法与 DPM 推理过程

收稿日期：2018-11-17；**修回日期：**2019-01-09 **基金项目：**河南省高等学校重点科研项目（19A520019）；山西省基础研究计划项目—青年科技研究基金（2014021039-6）

作者简介：孙新领（1981-），男，河南项城人，讲师，硕士，主要研究方向为计算机视觉、图形图像处理；张皓（1983-），男，河南新乡人，工程师，硕士，主要研究方向为图像处理、机器学习；赵丽（1980-），女，山西长治人，副教授，硕士，主要研究方向为图像处理、计算机应用等.

中的卷积操作完全相同，而 PS 作为部件模型与 DPM 有着相似的结构，该模型将部件划分为多个子类型，通过子类型的搭配组合可表示数目庞大的姿态形式。根据人体部位和人体部位类型设计 PS 一元滤波器，类似于 CNN[11]中的卷积层一样，该 PS 过滤器定义了外观模型的权重，因为它给出了特征相似性分数。该方法可以同时学习结构模型和外观模型，然后反向传播误差以学习底层的可学习参数。最终的对比实验也证明了提出的基于结构化 SVM 卷积神经网络的层次化人体姿态估计方法具有较强的识别性能。

1　模型和检测推理

1.1　结构化支持向量机（structured support vector machines, SSVM）两层神经网络

为了解决每个人体部位可能存在的不同外观，设计时使每个人体部位模型都包含多个不同的部位类型。从训练图像获得身体部位，根据它们在相对于相邻关节的图像坐标中的相对关节位置，把它们聚类成部位类型。这种聚类方法的基本假设是，同一组相关关节的位置外观上将很相似。共现模型考虑了在什么情况下，根据偏差系统两个相邻部分会共同出现[12,13]。相邻节点的每种混合类型都有相关的偏差。在众所周知的图像结构模型中结合这些措施，其边缘是根据以下假设进行量化的：假设放置部位所需的能量仅基于相对距离的二次变化，例如，从相对父节点的锚点位置拉伸或压缩弹簧所需的能量。根据文献[8]提出的方法，直接开发出了这些模型。他们将这三个模型，即共现模型、可变形模型和外观模型，组合成一个单一的大模型。在本文的研究中把前两种模型称为结构模型。

图 1 展示了提出的 SSVM 两层神经网络，其中第一层是卷积层（图 1 中的 SSVM-PS 层），另一层是损失增强推理层（如图 1 中的损失增强推理层）。通过将模型的结构化形式转换为模型中的一个神经网络，使其可以同时学习结构模型和外观模型，然后通过反向传播误差以学习底层的可学习参数，这些参数可从外观模型特征中提取出来（图 1 中的 CNN 层）。本文提出的方法将 SSVM 模型转换为神经网络模型，因此它具有神经网络的固有能力，将误差反向传播到较低层，并计算确切的 SSVM 损失，同时通过基于次梯度的方法来学习原始 SSVM。

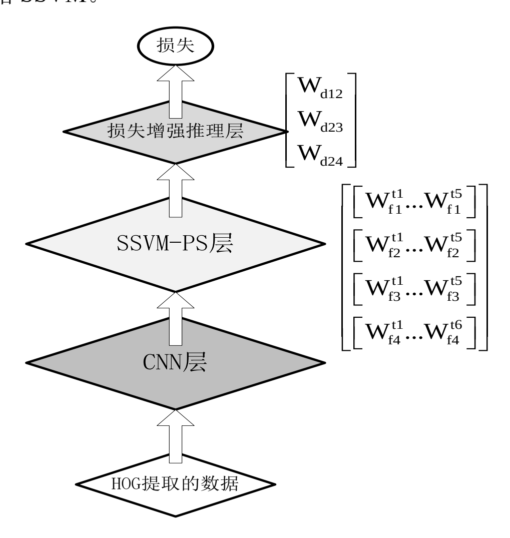


图 1　提出的 SSVM 两层神经网络

Fig. 1　Presented by the SSVM two-layer neural network

1.2　基于约束的 PS 模型

PS 模型将人体描述成一个无向图，无向图的边表示运动学上相连的两个部件，而每个节点表示一个身体部件。通常用矩形来表示节点：$l=(x,y,\theta,s)$。其中，(x,y) 表示部件的位置，θ 表示部件的方向，s 表示部件的尺度。则人体的姿态可定义成 $L=(l_1,l_2,\ldots,l_n)$。PS 模型方法是根据对身体各个部件之间的关系进行建模[14,15]。本文使用的是基于树型结构的 PS 模型，如图 2 所示，将人体的上半身分成头部、躯干、右上臂、右下臂、左上臂、左下臂。

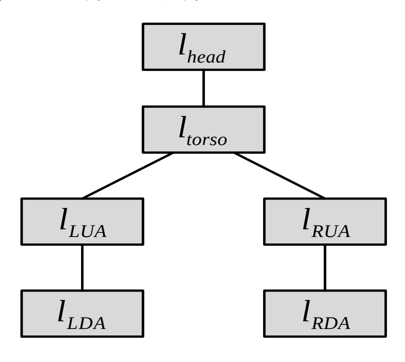


图 2　上半身 PS 模型

Fig. 2　Upper body PS model

假设人体的各个部件之间是相互独立的，设 L 代表人体各部件位置信息，I 代表的是图像信息，D 代表人体结构模型参数集。估计某一幅图像 I 中人体的姿态 L，根据 Bayes 理论，其后验概率可以表示为[16,17]

$$P(L/I,D)\propto\exp\left(\sum_{(x,j)\in E}\psi(l_i,l_j)+\sum_i\phi(I/l_i,\mathrm{D})\right)\tag{1}$$

其中：$P(L/I,D)$ 表示当模型为 D、图像 I 的情况下，人体的姿态是 L 的后验概率，$\phi(I/l_i,\mathrm{D})$ 代表外观模型上的部件 i 和在特定的位置 l_i 的图像特征的似然程度。而二元约束项 $\psi(l_i,l_j)$ 表示运动学上相连的两个部件 i 和 j 的位置的先验概率。

本文使用了一种约束的 PS 模型，如图 3 所示，它增加 $\gamma(l_{head})$ 和 $\gamma(l_{torso})$ 来限制躯干和头部的方向是竖直的，并通过给式(1)增加约束条件来实现。这是因为通常会遇到只有上半身可见的图像，而在这时，通常是假设人体的头部处于躯干之上。

$$p(L/\mathrm{I})\propto\exp\left(\sum_{(i,j)\in E}\psi(l_i,l_j)+\sum_i\phi(l_i)+\gamma(l_{head})+\gamma(l_{torso})\right)\tag{2}$$

为了减少头部和躯干的搜索空间，提高正确估计人体姿态的可能性，式中 $\gamma(\cdot)$ 表示在竖直方向附近的 θ 值概率是均匀的，而在其他方向上的概率为零。此外，为了提高上、下手臂的姿态估计准确率，通常情况下，会在确定躯干位置后，根据运动学上的先验概率 ψ 来限制手臂动作。

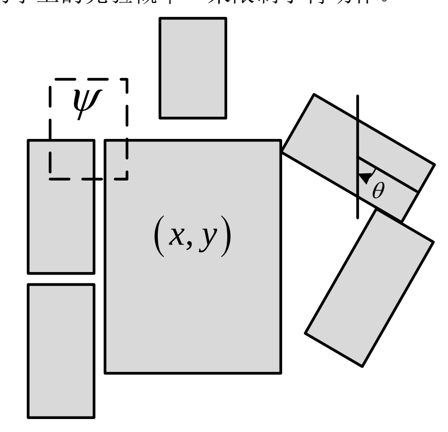


图 3　基于约束的 PS 模型

Fig. 3　Constraint-based PS model

1.3　子模型

将共现模型、可变形模型和外观模型组合成一个大模型，这三个子模型定义如下：

a)外观模型。外观模型包含各个部位的单独过滤器，如头部滤波器和身体滤波器。图像表示通常具有多个通道，因此这些模型都由包含滤波器大小乘以通道数量的矩阵表示。对于 5×5，32 或 64 通道的滤波器，其典型值分别为 5×5×32 或 5×5×64。通过确定滤波器的点积和相同大小的特征，可以获得特征的特定分数。这些滤波器位于 $R^{S\times C}$ 域中，其中 S 是滤波器大小，C 是通道数。对于每个部位和部位的每种类型，都有一个关联的外观模型滤波器。由滤波器 w_f^t 创建的相似度分数为

$$score_{appearance}(y)=w_f^t\cdot\Phi_f(x_L,y) \tag{3}$$

b)共现模型。假设某一部位有 m 个混合类型，而相邻部位有 n 个混合类型，那么这两个部位之间的总偏差为 $m\times n$。这个模型给出了局部得分和成对得分的总和。对于父节点 i 和子节点 j，共现分数 ij 为

$$score_{coocurrence}(t_i,t_j)=b_{ij}^{t_it_j} \tag{4}$$

可以把这视为偏向一些特定局部类型的偏差，以及父类和子类之间的配对关系。例如，如果 $b_{34}^{t_1t_2}$ 的值较高，这意味着父类部位编号为 3 的类型 1 可能会连接到子类部位编号为 4 的类型 2。

c)可变形模型。从每个父类 t_i 到每个子类 t_j，都有子类对父类的定位位置，其中从父类到子类，总共有 $t_i\times t_j$ 个定位点（锚点）。训练锚点的位置，以便利用所有可能的连接类型对其进行简单的建模。在 SSVM 训练之前，锚点位置必须可用，因此对每种外观类型，使用简单 K-均值聚类来计算部位类型，以创建关节的不同类型。本文使用组件的混合来解决可能部位外观的许多类型。令 $p\in P$ 是第 p 个身体部位，其中 $P=\{1,\dots,p_n\}$ 是所有部位的集合。令 $k\in K$ 是特定部位的第 k 种类型，其中 $K=\{1,\dots,k_n\}$ 是特定部位所有类型的集合。令 K_p 表示特定部位 p 的类型 K 的总数。

首先将训练图像部位 $p\in P$ 聚类为 K_p 聚类。现在，定义第 i 个样本的 SSVM 特征函数。将一元特征 $\Phi_f(x_{iL},y_i)$ 定义为 $\Phi_f(x_{iL})$，以便在位置 y_i 处进行评估。定义成对特征为 $\psi_{ij}=-[dx_{ij}\quad dy_{ij}\quad dx_{ij}^2\quad dy_{ij}^2]$ 其中，$dx_{ij}=x_{pi}-x_{pj}+锚点_x$，$dy_{ij}=y_{pi}-y_{pj}+锚点_y$。本文将 Ψ 表示为 $[\psi_{ij}]$，$\forall ij\in E$。对树 G 的一元特征和成对特征进行积分，得到 $\Phi_a=[\Phi_f\quad\Psi]$，其中下标 a 表示所有总和。由边缘 ij 的可变形模型获得的分数为

$$score_{defrom}(i,j)=W_{ij_d}\cdot\psi_{ij} \tag{5}$$

1.4　组合子模型

对于图形结构中的每个节点和边缘，本文将所有偏差权重、可变形权重和外观滤波器权重连接成数据结构的两种类型。第一种是基于组件的结构类型，第二种是向量类型。通过使用向量类型数据结构，创建了一个包含可学习 HPE 参数的大向量 W，用于 SSVM 学习。

$$score(t,y)=\sum_{i\in V}w_f^t\cdot\Phi_f(x_L,y)+\sum_{ij\in E}b_{ij}^{t_it_j}\cdot W_{ij_d}\cdot\psi_{ij} \tag{6}$$

或以矩阵形式表示为

$$score(t,y)=W\cdot\Phi_a(x_L,y) \tag{7}$$

为了找到分数值最大的位置 y，上面的等式变为

$$\hat{y}=\arg\max_{\hat{y}\in Y}W\cdot\Phi_a(x_L,y) \tag{8}$$

这是预测函数。$\hat{y}$ 的值是对一个测试图像中部位位置的预测。

1.5　SSVM 的次梯度优化

SSVM 的目标是通过学习每个训练数据的最大边缘分类器，来产生结构预测。像马尔可夫随机场（Markov random filed，MRF）或条件随机场这样的概率图模型，可以在学习阶段使用 SSVM 来学习权重参数[18,19]。SSVM 这种算法和 SVM 不一样，SVM 可以简单地插入数据以进行学习或分类，相反，SSVM 是一种在使用前需要指定推理、损失和特征模块的框架。例如，如果将 SSVM 应用于 MRF，则必须指定 SSVM 将学习的 MRF 结构和 MRF 推理算法，以及 MRF 特征函数、损失函数和损失增强推理算法。损失增强推理算法是具有损失函数的推理算法[20]。接下来，SSVM 根据训练数据学习使预测最大化的权重。SSVM 学习结构预测函数，如式（9）所示。

$$\begin{aligned}&\min_{w,\xi}\frac{\lambda}{2}\|w\|^2+\sum_{i=1}^{l}\xi_i\\&s.t.\forall i,W\cdot\Phi_{ai}(x_i,y_i)+\xi_i\\&\geq\max_{y\in Y}W\cdot\Phi_{ai}(x_i,y)+\Delta(y,y_i)\end{aligned} \tag{9}$$

通过把上述目标函数最小化，本文可以学习使预测函数等式（8）的训练精度最大化的参数 W。SSVM 目标函数式（9）可以通过次梯度方法求解，目标损失函数的次梯度定义为

$$\frac{\partial Obj_i}{\partial W}=\Phi_{ai}(x_i,\hat{y}_i)-\Phi_{ai}(x_i,y_i) \tag{10}$$

其中：Obj_i 是最小化目标函数式（9）。对于训练数据的第 b 批，梯度是

$$\frac{\partial Obj_b}{\partial W}=\frac{1}{b}\sum_b\Phi_{ab}(x_b,\hat{y}_b)-\Phi_{ab}(x_b,y_b) \tag{11}$$

其中：$\hat{y}_b$ 是根据损失增强推理得到的最违反的约束条件。

在本文中，本文把损失增强推理定义为

$$\begin{aligned}\hat{y}=&\arg\max_{y\in Y}\Delta(y,y_i)+W\cdot\Phi_a(x_i,y)\\&-W\cdot\Phi_a(x_i,y_i)\end{aligned} \tag{12}$$

其中：$\Delta(y,y_i)=1-\dfrac{Area(y\cap y_i)}{Area(y\cup y_i)}$ 是标准 1 减去联合损失中的边界框交集。然后，应用正常的随机梯度下降来进行梯度更新。

2　损失增强推理函数的求解

本文通过反向传播进一步向下传递到神经网络的较低层，扩展了先前定义的 SSVM 次梯度优化，这是因为 SSVM 可以实现为两层神经网络。顶层是损失增强推理层，底层是神经网络中正常的线性层。在将 PS 作为 CNN 的特殊情况下，底层是标准卷积层。

为了通过次梯度优化来求解 SSVM，必须计算损失增强推理 $\hat{y}_i$，以便可以计算式（10）中最违反约束的特征 $\Phi_{ai}(x_i,\hat{y}_i)$。注意到，w_f^t 对 $\hat{y}$ 可能位置上的所有 $\Phi_f(x_{iL})$ 的滑动点积，实际上是一个卷积运算，完全等于 CNN 中的卷积层，相应的前馈是

$$\Phi_{rf}^t(x,y)=\sum_{\bar{y}}W_f^t(\bar{y})\Phi_a(x,y+\bar{y}) \tag{13}$$

因此，式（13）将 SSVM 的底层定义为两层神经网络。实际上，数量是 $\Phi_f(x_{iL})$ 与 PS 一元滤波器 W_f^t 卷积的响应映射。这里，把 Φ_r 表示为所有 Φ_{rf}^t 的级联。使用构造的 SSVM-PS

层，并在这个两层神经网络环境中定义损失增强推理层。显然，损失增强推理层执行损失增强推理并寻找目标函数的松弛损失，$\hat{y}$ 是使损失增强推理目标函数最大化的最违反约束参数。可以使用式（6）（8）和（13）把损失增强推理目标函数式（12）重写为

$$\begin{aligned}L_i=&\max_{y\in Y}\Delta(y,y_i)+\sum_{v\in V}\{\Phi_r(x_{iL},y_v)-\Phi_r(x_{iL},y_{vi})\}\\&+\sum_{ef\in E}\{W_{ef_d}\cdot\Psi_{ef}(x_i,y)-W_{ef_d}\cdot\Psi_{ef}(x_i,y_i)\}\\&+\sum_{ef\in E}\{(b_{ef}^{t_e t_f})_y-(b_{ef}^{t_e t_f})_{y_i}\}\end{aligned}\tag{14}$$

在神经网络意义上，可以把 $\Phi_r(x_{iL},y)-\Phi_r(x_{iL},y_i)$ 这一项视为从下层矩阵中选取两个标量并执行减法。因此，式（14）将 SSVM 的上层定义为两层神经网络，可以看到式（8）中的 W 现在被分成两个不同的层。外观模型方程式（3）的权重位于卷积层的底部。共现模型和可变形模型式（4）（5）的权重位于损失增强推理层的顶部。

接下来，定义两层神经网络的反向传播规则。顶层（即损失增强推理层）的梯度是

$$\begin{aligned}\frac{\partial L_b}{\partial W}=&\frac{1}{b}\sum_b\{\Psi_b(x_b,\hat{y}_b)-\Psi_b(x_b,y_b)\\&+[\delta(\hat{y}^{t_i t_j})]_b-[\delta(y_i^{t_i t_j})]_b\}\end{aligned}\tag{15}$$

其中：$[\delta(a)]$ 是一个向量，其元素除了 a 处位置的元素之外全部为零。目标函数对响应映射层 $\Phi_r(x_{iL})$ 的梯度为

$$\frac{\partial L_i}{\partial \Phi_r}=\delta(y_v,\hat{y}_v)-\delta(y_v,y_{iv}),\forall y\in Y,\forall v\in V\tag{16}$$

其中：如果 $a=b$，则 $\delta(a,b)=1$，否则 $\delta(a,b)=0$。这可以通过下面的方法验证：创建与 $\Phi_r(x,y)$ 同大小的空矩阵，然后令 $\bar{y}_v$ 位置为+1，令 y_{iv} 位置为-1，但是如果存在 $\bar{y}_v$=y_{iv} 的点，则令该位置的值为 0。上面指定的两个梯度定义了损失增强推理层。根据SSVM-PS层的反向传播规则来定义图形结构的外观层，本文可以对 CNN 的卷积层使用正常的反向传播规则。

要用最大化式（14）的方法求解 $\hat{y}$，本文使用标准最大和（max-sum）算法。最大和算法的目的是通过以下形式，在图 $G=\{V,E\}$ 的情况下找到组合优化问题的解。

$$\hat{L}=\arg\max_L\sum_{i\in V}m_i(l_i)+\sum_{ij\in E}g(l_i,l_j)\tag{17}$$

组合优化问题式（14）的目标函数是

$$\begin{aligned}\hat{y}=&\arg\max_{y\in Y}\sum_{v\in V}\{\Phi_r(x_{iL},y_v)+\Delta(y_v,y_{iv})\}\\&+\sum_{ef\in E}\{W_{ef_d}\cdot\Psi_{ef}(x_i,y)+(b_{ef}^{t_e t_f})_y\}\end{aligned}\tag{18}$$

在神经网络的最顶层执行最大和算法，来求解这个组合目标问题。

3 实验

与训练神经网络的方式相同，本文用随机梯度下降训练 SSVM。本文网络架构是将正常的 CNN 与 SSVM 神经网络相连接，作为最后两层。在深度学习框架(convolution architecture for feature extraction, Caffe)中，把损失增强推理作为一层实现。

3.1 数据准备

本文采用了 PARSE 数据集，包括 100 个正面训练样本和 205 个正面测试样本。该数据集中的每张图像都显示了人的整个身体，通常是在运动环境中。对于每个样本，都标记了相应的人体关节位置。在图像中，一些人体部位被遮挡，但提供了人体关节位置的估计。每个样本总共标注了 14 个关节，包括头部、躯干、左臂、右臂、左腿和右腿。图像的大小范围为[110-450]×[110-450]。在一些图像中存在一个或两个完整的人体。图像中的人体尺寸也有所不同。数据集中的人体姿态变化：从坐着到站着，腿部和手臂可能会做着诸如武术或体操等运动。

该数据集不适合直接在本文算法中使用，因此对其进行了预处理，如下所示。把训练数据与训练标签镜像翻转，然后添加到原始训练数据中。因此，本文有一组双倍大小的训练数据。然后，找到每两个关节的中间点，从而总共获得 26 个关节点和中间关节点。本文将这些点改变为方框，其中方框大小是通过对训练数据中关节的长度取平均值来计算的。在这个阶段，对每个方框，把本文训练算法中标签定义为 (x_1,y_1,x_2,y_2)。每个训练标签有 26 个方框，包含 y 个图像空间。然后，使训练图像通过 HOG 金字塔特征。在这个过程中，基于同一张训练图像，调整其大小以获得具有不同尺寸的多张图像，这被称为图像金字塔。使用 HOG 提取图像金字塔，以获得特征金字塔，然后将其填充到零矩阵中以获得 $\Phi_h(x_{iL})$。把标签添加到特征金字塔参数中，以生成实际标签。然后，将这个实际标签和批处理数据转换成内存映射数据库，这样，Caffe GPU 库可以更有效地处理这些数据。

3.2 神经网络结构

本文神经网络结构如下：数据层—卷积层 CnnFeat—卷积层 SSVM-PS—损失增强推理层。将卷积层置于中间，以此可以在中间学习深度学习特征提取参数。另外，权重是随机初始化的。本文混合模型中每个节点类型如下：

$$M_{ixture}=\{5,5,5,6,6,6,6,5,5,5,5,5,5,5,5,6,6,6,6,5,5,5,5,5,5,5\}$$

这意味着，第一个节点（头节点）有五种不同的混合类型，第二个节点有五种不同的混合类型，依此类推。总共有 $\sum_i M_{ixture}=138$ 种混合类型。SSVM-PS 层的大小表示为 5×5×256×138=883200。该损失增强推理层有 $1+\sum_i M_{ixture}(i)\times M_{ixture}(i-1)=702$ 个权重元素，以及 133×4=532 个 w_{deform} 权重元素。因此，损失增强推理层总共有 1234 个权重元素。

SSVM-PS 层共有 5×5×256×138=883200 个权重元素。本文将 CnnFeat 的内核大小设置为 2×2×32，因此 CnnFeat 总共有 2×2×32×256 = 32768 个权重元素。因此，系统总共有 883200+1234+32768=917202 个权重元素。设置各批次大小为 50，特征尺寸 Φ_h 为 140×140。数据层有 50×32×140×140=31360000 个元素。CnnFeat 层有 50×256×139×139=247308800 个元素。SSVM-PS 层有 50×138×135×135=125752500 个元素。总共需要 404421300 个单元来存储本文多层神经网络数据。需求的 GPU 总内存为 1284714404 字节。

本文以 0.005 的学习率训练了超过 3000 次迭代。使用 L_2 正则化项，系数为 0.1，没有动量参数。

3.3 Caffe 层实现

在数据层中的 $\Phi_h(x_{iL})$ 通过 CnnFeat 层之后获得 $\Phi_f(x_{iL})$，进一步向前通过 SSVM-PS 层产生 $\Phi_r(x_{iL})$ 作为响应映射或热映射。损失增强推理层使用 Caffe 库中的 python 层实现。损失增强推理层使用最大和算法来计算目标函数的值。最大和算法的输出包含：最优水平和最违反约束 $\hat{y}$，其输出特征值分别为 $\Phi_a(x_{iL},\hat{y})$ 和 $\Phi_a(x_{iL},y_i)$。通过搜索每个金字塔等级，然后找到最佳最大边际得分，从而得到最大边际值最大化方程式（18）以及损失函数 $\Delta(\hat{y},y_i)$。用新金字塔等级中的较高分数取代以前的结果。

3.4 结果

基于 PARSE 数据集，本文训练和测试了提出的方法。根据式（12）的损失增强推理损失 L_i。将本文的结果与文献[6~8]得到的结果进行了比较，如图 4 所示。正确检测标准为 PCP（percentage of correct parts）[8]，如果检测到的肢体端点和地面肢体端点之间的距离在肢体长度的一半之内，则认为肢体被正确检测到。本文方法在头部、躯干、左臂、右臂、左腿以及右腿共六个部位的姿态估计性能均高于其他几种对比方法。此外，将本文方法与文献[6~8]方法对 PARSE 数据集中的图像 Im0001-Im0008 图像（图 5）进行目标检测，结果如表 1 所示，类似于图 5 结果，本文方法对 8 幅图像的识别成功率也均高于对比方法。im0001-im0008 是 PARSE 数据集中的前 8 个图像。识别成功率是平均 pcp，这个数据是可以量化的。其中，文献[6]提出的带有附加潜变量的图形结构树模型以及文献[7]提出的多模态可分解模型，由于缺乏较好的先验模型，导致了其识别性能受到了较大限制。而文献[8]方法不能学习由深度学习特征而提取的参数。而且由可学习的特征所提取的参数不能基于 Latent SVM 的误差来更新。本文提出基于结构化 SVM 卷积神经网络的层次化模型将 SSVM 模型转换为神经网络模型，因此它具有神经网络的固有能力，将误差反向传播到较低层，并计算确切的 SSVM 损失，同时通过基于次梯度的方法来学习原始 SSVM 解决了这个问题。因而，对比结果均显示提出的方法具有更强的识别性能。

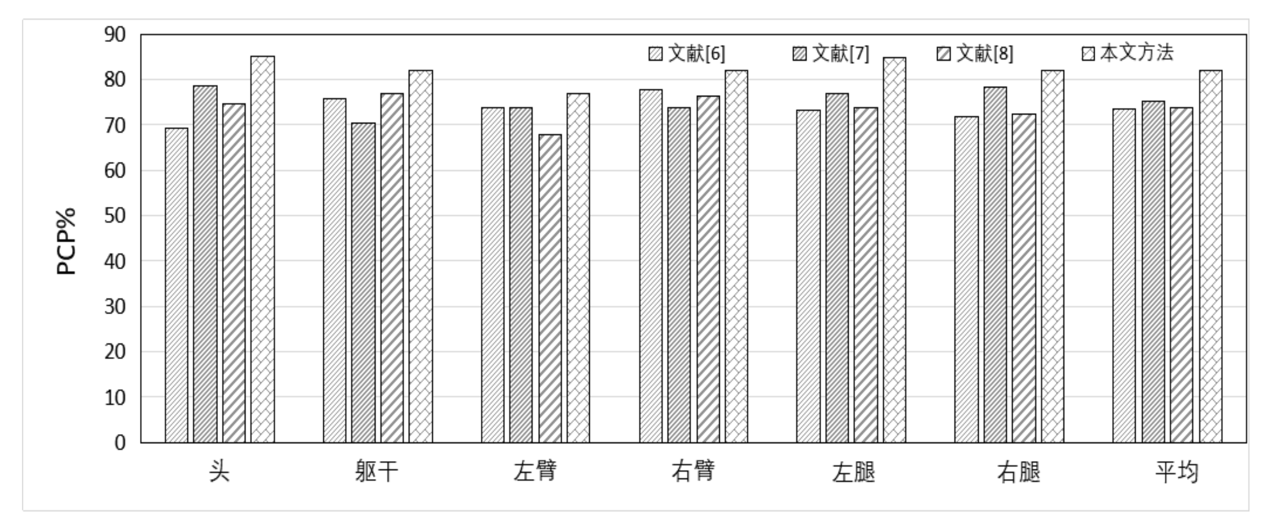


图 4　几种方法姿态估计性能对比

Fig. 4　The comparison of attitude estimation performance of several methods










im0001　im0002　im0003　im0004　im0005　im0006　im0007　im0008

图 5　PARSE 数据集中的图像 Im0001-Im0008

Fig. 5　Image Im0001-Im0008 in the PARSE dataset

表 1　几种方法目标检测性能对比

Table 1　The comparison of target detection performance of several methods

Model	识别成功率（平均 PCP）							
	Im0001	Im0002	Im0003	Im0004	Im0005	Im0006	Im0007	Im0008
文献[6]	66.89	76.95	68.95	68.94	73.53	75.34	78.34	68.24
文献[7]	68.94	65.97	76.45	78.65	72.87	69.85	79.43	78.56
文献[8]	65.97	73.95	76.46	67.43	80.23	70.54	76.23	77.67
本文方法	77.76	79.98	82.85	80.34	82.73	80.43	84.83	78.76

4　结束语

目前，许多基于部位的检测方法依靠 CNN 作为前端。许多研究表明，通过将分类器反向传播到深度学习特征提取器，可以获得更好的分类和特征提取性能。本文提出了一种具有深度 CNN 的新型 SSVM 神经网络层来解决基于部位的图像检测问题。实验结果证明，通过减少 SSVM 神经网络的损失，可以将这种方法很好地应用于基于部位的检测。

在未来的研究中，会创建一个完整的端到端神经网络，来解决 HPE 问题以及其他基于部位的检测问题。

参考文献：

[1] 刘兴旺，王江晴，徐科．一种融合 AutoEncoder 与 CNN 的混合算法用于图像特征提取 [J]．计算机应用研究，2017，34(12): 3839-3843. (Liu Xingwang, Wang Jingqing, Xu Ke. Novel image feature extraction algorithm based on fusion AutoEncoder and CNN [J]. Application Research of Computers, 2017, 34 (12): 3839-3843.)

[2] 吴素雯，战荫伟．基于选择性搜索和卷积神经网络的人脸检测 [J]．计算机应用研究，2017, 34(9): 2854-2857. (Wu Suwen, Zhan Yinwei. Face detection based on selective search and Gabor optimizing convolutional neural network [J]. Application Research of Computers,

2017, 34(9): 2854-2857.)

[3] Du Xianzhi, El-Khamy M, Lee J, *et al.* Fused DNN: a deep neural network fusion approach to fast and robust pedestrian detection [C]// Proc of IEEE Winter Conference on Applications of Computer Vision. Piscataway, NJ: IEEE Press, 2017: 953-961.

[4] Chan Kaichi, Koh C K, Lee C S G. An automatic design of factors in a human-pose estimation system using neural networks [J]. IEEE Trans on Systems Man & Cybernetics Systems, 2016, 46 (7): 875-887.

[5] Rafique M A, Azam M S, Jeon M, *et al.* Face-deidentification in images using Restricted Boltzmann Machines [C]//Proc of the 11th International Conference for Internet Technology and Secured Transactions. Piscataway, NJ: IEEE Press, 2016: 69-73..

[6] Steorts R C, Hall R, Fienberg S E. A Bayesian approach to graphical record linkage and deduplication [J]. Publications of the American Statistical Association, 2014, 111 (516): 1660-1672.

[7] Sapp B, Taskar B. MODEC: multimodal decomposable models for human pose estimation [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC:IEEE Computer Society, 2013: 3674-3681.

[8] Yang Songfan, Ramanan D. Multi-scale Recognition with DAG-CNNs [C]//Proc of IEEE International Conference on Computer Vision. Washington DC:IEEE Computer Society, 2015: 1215-1223.

[9] 宋瑨，王世峰. 基于可变形部件模型 HOG 特征的人形目标检测 [J]. 应用光学，2016, 37(3): 380-384. (Song Jin, Wang Shifeng. Human-kind shape object detection using deformable parts model with HOG features [J]. Journal of Applied Optics, 2016, 37(3): 380-384.)

[10] 李春伟，于洪涛，李邵梅，等. 一种基于可变形部件模型的快速对象检测算法 [J]. 电子与信息学报，2016, 38(11): 2864-2870. (Li Chunwei, Yu Hongwei, Li Shaomei, *et al.* Rapid object detection algorithm based on deformable part models [J]. Journal of Electronics & Information Technology, 2016, 38(11): 2864-2870.)

[11] 周飞燕，金林鹏，董军. 卷积神经网络研究综述 [J]. 计算机学报，2017, 40(6): 1229-1251. (Zhou Feiyan, Jin Linpeng, Dong Jun. Review of Convolutional Neural Network [J]. Chinese Journal of Computers, 2017, 40(6): 1229-1251.)

[12] Marangoni M N, Brady S T, Chowdhury S A, *et al.* The co-occurrence of myocardial dysfunction and peripheral insensate neuropathy in a streptozotocin-induced rat model of diabetes [J]. Cardiovascular Diabetology, 2014, 13(1): 1-10.

[13] Akbari M, Samadzadegan F, Weibel R. A generic regional spatio-temporal co-occurrence pattern mining model: a case study for air pollution [J]. Journal of Geographical Systems, 2015, 17(3): 249-274.

[14] Navarrocompán V, Gherghe A M, Smolen J S, *et al.* Relationship between disease activity indices and their individual components and radiographic progression in RA: a systematic literature review[J]. Rheumatology, 2015, 54 (6): 994-1007.

[15] Toderi S, Gaggia A, Mariani M G, *et al.* Griffin and Neal's safety model: determinants and components of individual safety performance in the Italian context. [J]. La Medicina Del Lavoro, 2015, 106 (6): 447-459.

[16] 董峰辉，程进. 极值-I 型风速预测的 Bayes 方法 [J]. 哈尔滨工业大学学报，2017, 49(3): 93-97. (Dong Fenghui, Cheng Jin. Bayesian method for extreme value-I wind speed prediction [J]. Journal of Harbin Institute of Technology, 2017, 49(3): 93-97.)

[17] Font O, Frances G, Jonsson A, *et al.* Probabilistic activity recognition in navigation [C]// Proc of the 11th Workshop on Positioning, Navigation and Communication. Piscataway, NJ: IEEE Press, 2014: 1-6..

[18] 汤浩，何楚. 全卷积网络结合改进的条件随机场-循环神经网络用于 SAR 图像场景分类 [J]. 计算机应用，2016, 36(12): 3436-3441. (Tang Hao, He Chu. SAR image scene classification with fully convolutional network and modified conditional random field-recurrent neural network [J]. Journal of Computer Applications, 2016, 36(12): 3436-3441.)

[19] 蔡波. 基于概率图模型的目标跟踪算法研究 [D]. 南京：南京航空航天大学，2016. (Cai Bo. Research on Target Tracking Algorithm Based on Probability Graph Model [D]. Nanjing：Nanjing University of Aeronautics and Astronautics, 2016.)